科学技術文献を対象とした日中機械翻訳システムの開発

日中・中日言語処理技術の開発研究

豊橋技術科学大学情報メディア基盤センター教授 **井佐原　均**

PROFILE

電子技術総合研究所（現、産業技術総合研究所）、情報通信研究機構を経て、現職。アジア太平洋機械翻訳協会会長。言語資源協会理事。

1 はじめに

科学技術の健全な発展のためには、特許を含む科学技術情報の国際的な流通が不可欠である。特にアジア圏においては、科学技術面で飛躍的な発展を遂げつつある中国をはじめ、アジア諸国内において流通している科学技術情報の日本国内での活用を容易にし、かつ我が国が最先端を担っている科学技術分野の文献情報の各国への流通を促進することが重要である。アジア諸国と日本の科学技術の発展に資することを目指し、まず中国語を対象に科学技術論文を対象とした機械翻訳システム、および関連する技術、利用する言語資源を科学技術振興調整費の助成を得て、平成 18 年度から 5 年間で開発した。

このプロジェクトは、独立行政法人情報通信研究機構を取りまとめ機関とし、独立行政法人科学技術振興機構、東京大学、京都大学、静岡大学の参画を得て、実施された。具体的には、用例ベース翻訳の研究開発を京都大学と情報通信研究機構が、言語資源の構築の研究開発を東京大学、静岡大学、科学技術振興機構が、それぞれ主として担当した。

用例ベース翻訳の研究開発においては、論文等の定型性のある文書に有効である用例ベース翻訳手法において顕著な成果を挙げている京都大学と、解析システムの経験が豊富な情報通信研究機構の組み合わせは、特に翻訳精度の向上において中国語解析の精度向上が必須であったことから、有効であった。

また、言語資源の構築の研究開発においては、用語抽出と辞書構築の成果を持つ東京大学、語義ネットワークの先進的研究を行う静岡大学を大規模な論文データベースを持つ科学技術振興機構の言語資源が底支えするという体制であり、これも有効な研究体制であった。なお、ほとんどの組織が研究参加者として中国人の研究者を雇用し、中国語の解析・翻訳・言語資源の開発に必要な知見を得る手段を確保していた。

本稿では、このプロジェクトの成果を概観する。

2 研究計画とミッションステートメント

日本語と中国語のように言語構造の異なる言語間の翻訳に適し、かつ科学技術文書の概要を把握することに利用できる高精度の翻訳システム、および高度な専門用語や新出用語にも対応できるような大規模な機械翻訳用辞書を研究開発するため、以下の項目において目標設定を行い、研究開発を実施した。

①日中・中日の用例ベース翻訳のための要素技術の研究開発（サブテーマ 1）

○解析システムに関する研究開発

これまで英語や日本語の解析で有効性が確認されてきたコーパスに基づく解析手法を中国語に適用し、中国語の形態素解析・構文解析の高度化を図る。

○翻訳エンジンに関する研究開発

用例ベース翻訳において用例を柔軟に利用できるようにするとともに、対訳コーパスにおいて語・句を高精度で対応付ける手法を確立する。

②日中・中日言語資源の構築と構築技術に関する研究開発（サブテーマ 2）

○日中・中日翻訳用大規模辞書の構築

・日英（及び英中）の専門用語・一般用語対訳辞書をもとに初期版の日中・中日辞書を作成し、さらに当該辞書を活用し、基本用語と科学技術用語（同義語、異表記語を含む）の半自動収集を行い、対訳関係を収集することで機械翻訳のための大規模辞書を作成する。

・上記辞書を用いて高精度な文解析、翻訳処理に必要な意味関係の抽出、および多義語・多訳語に対処するために必要な個別の語ごとの統計的なプロファイルを持った辞書を構築する。

・分野依存的に関連専門用語の使用パターンを統計的に分析することで、複数の用語間に現れる因果関係などのより豊かな意味関係を抽出し辞書に付加する手法を確立する。

・文法規則解析用に、文献データベース中に存在する中国文献のタイトル、抄録および教科書などの各種の科学技術関連文書より、大規模な日中英文献コーパスを作成し、このコーパスから科学技術文献対応の語義関連ネットワークを生成し、得られた語義関連ネットワークと訳語選択プログラムを用いて、評価例文に対する訳語選択を実行し訳語選択の精度を評価する。

○日中・中日翻訳用大規模対訳コーパスの構築

日中・中日の大規模なコーパスを収集するとともに、バランスのとれた収集方法や、対訳文対を効率良く増やす手法を確立する。

③日中・中日機械翻訳プロトタイプシステムの開発および実証実験（サブテーマ３）

○上記で得られた研究成果をもとに、日中・中日機械翻訳プロトタイプシステムを作成し、実用レベルに近い機械翻訳が実現可能であることを示す。

このような計画の下、以下の２段階のミッションステートメントを決定した。

A．３年の中間段階におけるステートメント

○対訳コーパス（100 万文規模）の作成（データの収集、翻訳）と半自動解析

○辞書（日英中専門用語辞書、格フレーム辞書、意味体系）の作成

○辞書半自動構築システムの構築と評価

○解析・翻訳エンジンの開発・改良

○特定の対象分野について、日中機械翻訳プロトタイプ実証システムの動作を確認し、情報通信研究機構のサイトなどで日中機械翻訳を体験できるようにする。

B．研究終了段階におけるステートメント

○日本語および中国語の科学技術文献を対象に、翻訳率 80%以上を実現する高品質な日中・中日機械翻訳プロトタイプシステムを開発する。

○科学技術文献検索・翻訳の試行的なサービスを行い、その有用性を検証・評価する。

○実施期間中随時および終了後速やかに、著作権をクリアした上で他者が使用できる形態で、成果となる対訳コーパス、翻訳用辞書、翻訳エンジン、および評価に使用したデータセットを公開する。

このような実施体制、研究計画、ミッションステートメントの下、研究開発を実施した。３年目および研究終了段階でのミッションステートメントを満たすことが出来たことから、研究計画は基本的に適切なものであったと判断している。なお、研究の途上で、用例ベース翻訳システムに必須な対訳データベースの開発に想定以上の費用がかかることが分かったり、中国語に関する翻訳の精度向上には中国語解析の精度を日本語や英語と同水準に上げる必要があることが分かったりしたが、予算や人的配置を適切に組み替えることにより、当初計画どおりの研究進捗が可能となった。

3 研究成果

平成20年６月中旬までの約２年間の研究開発は種々の問題を解決しつつ順調に進み、３つのサブテーマに関して当初の中間目標を３年間で十分に達成できる見通しが得られた。これらの成果を元に中間評価を受け、ア

ドバイスを頂いた。プロジェクトの後半では、残りの課題を克服しつつ計画通り進めた。

サブテーマ1の「日中・中日の用例ベース翻訳のための要素技術の研究開発」においては、用例ベース翻訳エンジンの研究開発を進め、既存データ100万文が活用できる日英翻訳において標準的な統計翻訳エンジンと同等以上の性能を達成し、科学技術文献の翻訳システムとして最先端の性能を実現した。これと並行して、中国語解析など基本技術の研究開発も進め、中国語解析に関しては世界最高の性能に達した。また、サブテーマ2で作成した辞書と対訳コーパスを活用し、辞書の追加により翻訳性能が向上することを確認した。

日中翻訳における具体的な翻訳精度は、人間による主観評価で、日中方向については87.5%、中日方向については76.5%の翻訳率[1)]である。この値は商用システム（J-server）より高精度であり、統計翻訳システム（Moses）にはわずかに劣る。一方、英日・日英翻訳では、本手法が統計翻訳や商用システムよりも高性能である。これは、日本語や英語の解析システムの精度が、中国語解析よりも高いためと考えられる。

現状が商用システムより高精度であることから、今後の改良のためには、用例ベースなどの言語資源の開発が有効であることが分かる。また、英日・日英では統計翻訳より高精度であることから、中国語解析の改良が有効であることが分かる。

日中	人手評価 2以上	自動評価 BLEU
用例ベース	87.5%	27.1
J-server	74.5%	8.8
Moses	92.5%	35.2

日中翻訳実験結果

中日	人手評価 2以上	自動評価 BLEU
用例ベース	76.5%	35.5
J-server	34.8%	7.3
Moses	82.5%	46.2

中日翻訳実験結果

用例ベースの翻訳手法で最も重要なのが、対訳コーパスにおける語句の対応付けであり、この性能向上が（特に日中のように言語タイプが異なる場合には）重要である。本プロジェクトでは基本的な文法規則を使うことなどによって、統計的な手法に比べて高精度なアラインメント結果を得ることができた。これにより、高精度の翻訳と、質の高い用例辞書データベースの構築が可能となった。

また、中国語解析システムにおいては、形態素解析（単語区切りと品詞付与）では93.7%の精度を達成した。これは世界最高性能である。構文解析（依存構造解析）では91.9%の精度を達成し、これは標準的なデータを対象として、世界で初めて90%超の精度を達成したものである。

サブテーマ2の「日中・中日言語資源の構築と構築技術に関する研究開発」においては、目標の日中対訳コーパスの構築と半自動解析が3年間で概ね完了した。辞書については、日英と英中の既存の辞書を用いて中間言語として英語を介することにより対訳辞書を自動構築する方法、さらに自動構築の際に混入するノイズを除去する方法を考案・評価した。類義語を抽出する方法、格フレームを抽出する方法等も考案・評価し、各辞書を試作した。さらに、これらを統合したシステムの設計を行った。

用例ベース翻訳手法では、言語資源として大量の用例を蓄積する必要があり、本プロジェクトでは2700万字、45.6万文の日中対訳コーパス（日本語と中国語の対訳文）を作成した。原文は文献データベースの抄録と、電子ジャーナルJSTAGE登載の文献本文を選び、後者については学会の許諾を得て利用した。文例は文献抄録からが2割、文献本文からが8割である。

1）翻訳率：専門家にとって理解できる程度の訳文の割合。人手による3段階評価の上位2段階。

	文字数	文数	割合
医学	7,623,493	128,026	28%
環境	2,877,030	52,589	12%
化学	1,709,963	26,513	5%
農学	1,130,518	17,647	4%

	文字数	文数	割合
生物	4,163,659	65,957	14%
材料	1,360,440	23,406	5%
エネルギー	698,487	11,867	3%
情報	1,709,963	129,729	28%
計	27,053,801	455,734	

作成した日中対訳コーパス

サブテーマ３「日中・中日機械翻訳プロトタイプシステムの開発および実証実験」の目的は、サブテーマ１および２の成果を統合し、プロトタイプシステムを構築することである。プロトタイプシステムを構築し、日英・日中のそれぞれ双方向で動作を確認した。機械翻訳性能を自動評価するための新たな手法も考案した。

4 アウトリーチ活動

日本国内の関連分野の研究者を対象とする情報発信としては、研究の全容と各年度の進捗について、各年度末に開催した公開シンポジウムにより十分な情報発信を行った。研究内容が日本語と中国語の間での機械翻訳システムの開発であることから、特に中国の研究者に向けての情報発信を重視し、毎年、日中自然言語処理共同研究促進会議に本プロジェクトの全チームが参加し、研究成果を発表し、討論を行った。また、より広く世界中の機械翻訳研究者への情報発信として、平成 19 年にヨーロッパで開催された機械翻訳サミットにおいて、本プロジェクトの全容を報告した。

より一般向けの情報発信としては、国内においては情報通信研究機構の一般公開でデモ展示を行った。また中国においては、情報通信研究機構が展示を行うのにあわせて、デモ展示を行った。さらに一般大衆に向けての認知度を上げるため、北京五輪において、北京組織委員会の公式翻訳システムに組み込まれ、また日本人向けのジャパンハウスでの展示を行うなど、積極的な広報活動を行った。

5 これから

今回開発したシステムは実用に極めて近いシステムと考えているが、広範な実場面での利用には、いくつか超えなければならないハードルがある。今後は中国との国際協力や、公開データを用いた広い研究協力により、真に実用可能なシステムへと発展させていきたいと考えている。